Inferring Gene-Gene Associations from Quantitative Association Rules

M. Martínez-Ballesteros, I. Nepomuceno-Chamorro, J.C. Riquelme
Department of Computer Science
University of Seville
Seville, Spain
Email: mariamartinez,inepomuceno,riquelme@us.es

***Abstract*—The microarray technique is able to monitor the change in concentration of RNA in thousands of genes simultaneously. The interest in this technique has grown exponentially in recent years and the difficulties in analyzing data from such experiments, which are characterized by the high number of genes to be analyzed in relation to the low number of experiments or samples available. Microarray experiments are generating datasets that can help in reconstructing gene networks. One of the most important problems in network reconstruction is finding, for each gene in the network, which genes can affect it and how. Association Rules are an approach of unsupervised learning to relate attributes to each other. In this work we use Quantitative Association Rules in order to define interrelations between genes. These rules work with intervals on the attributes, without discretizing the data before and they are generated by a multi-objective evolutionary algorithm. In most cases the extracted rules confirm the existing knowledge about cell-cycle gene expression, while hitherto unknown relationships can be treated as new hypotheses.**

***Keywords*-Data mining; evolutionary algorithms;quantitative association rules; gene networks**

I. Introduction

Microarray technology has revolutionized the biological research due to its ability to monitor changes in RNA concentration in thousands of genes simultaneously [1]. Research in molecular biology has traditionally focused on the study gene to gene, but nowadays we are in the genomic era and genes are studied in thousands or even whole genomes. Standard approaches to microarray analysis (biomarker discovery) are based on the identification of differentially expressed genes and the assumption that genes act independently. However, it is known that powerful prognostic biomarkers may be encoded by genes that are not highly differentially expressed across control and disease patients [2]. Therefore, a systems-level approach can provide insights into the interplay of genes and their association with clinical phenotypes.

In this context we present the result of applying a data mining technique, specifically, association rules, to gene expression data from experiments using microarray technology. The aim of mining association rules is to discover the sets of attributes which appear in a dataset with a certain frequency in order to obtain rules that show the existing relationships among the attributes, specifically, this technique is applied to discover associations between genes from microarray datasets, in which gene expression is linked to another gene expression.

A revision of the published literature reveals that exist many algorithms such as Apriori [3] to find ARs. However, many of these tools that work in continuous domains just discretize the attributes by using a specific strategy and deal with these attributes as if they were discrete [4]. Many algorithms are based on evolutionary algorithms (EAs) [5] which have been extensively used for the optimization and adjustment of models in data mining tasks. EAs are used to discover ARs due to they offer a set of advantages for knowledge extraction and specifically for rule induction processes [6]. In [7] the authors proposed an EA to obtain numeric ARs, dividing the process in two phases. Another EA was used in [8] to obtain ARs where the confidence was optimized in the fitness function.

The mining process of ARs can be considered as a multi-objective problem rather than a single objective one, in which the measures used for evaluating a rule can be thought as different objectives. In the last two decades an increasing interest has been developed in the use of EAs for multi-objective optimization [9]. There are multiple proposals such as the algorithms NSGA II [10] or SPEA2 [11] for instance. In [12] a multi-objective pareto-based EA was presented and another multi-objective GA to AR mining is proposed in [13].

In preliminary works such as the proposed algorithms in [14] and [15], henceforth called QARGA (Quantitative Association Rules by Genetic Algorithm), authors of this paper developed several single-objective EA that use a weighting scheme for the fitness function which involved some evaluation measures. However, it is known that a scheme of this nature is not ideal compared to multi-objective schemes, so that could reduce the features used in the fitness function for applying a multi-objective technique.

Thus, the main motivation of this paper is to extend these algorithms to a multi-objective approach based on the NSGA-II algorithm. The non-dominated multi-objective evolutionary algorithm proposed in this work can find quantitative association rules in databases with continuous attributes from microarray data, avoiding the discretization as a step in the process. The results will show that the rules obtained have been able to successfully characterize the data underlying and also to group relevant genes for the problem studied.

The rest of the paper is organized as follows. Section II provides a brief preliminary on ARs. Section III describes the methodology used in this work. The results obtained by the developed algorithm are discussed in Section IV. Finally, Section V provides the achieved conclusions.

II. Association rules

Data mining is one of the most used instrumental tools for discovering knowledge from transactions. In the field of data mining, the learning of ARs is a popular and well-known research method for discovering interesting relations among variables in large databases [3]. The discovery of ARs is, unlike classification, a non-supervised learning tool as ARs are descriptive. Descriptive mining tasks identify patterns that explain or summarize the data, that is, they are used to explore the properties of the data, instead of predicting the class of new data [16].

This form of knowledge extraction is based on statistical techniques such as correlation analysis and variance. One of the most widely used algorithms is the Apriori algorithm.

Formally, AR were first defined by Agrawal et al. in [17] as follows. Let $I = \{i_1, i_2, ..., i_n\}$ be a set of n items and $D = \{t_1, t_2, ..., t_N\}$ a set of N transactions, where each t_j contains a subset of items. Thus, a rule can be defined as $X \Rightarrow Y$, where $X, Y \subseteq I$ and $X \cap Y = \emptyset$. Finally, X and Y are called antecedent (or left side of the rule) and consequent (or right side of the rule), respectively.

When the domain is continuous, the ARs are known as Quantitative Association Rules (QAR). In this context, let $F = \{F_1, ..., F_n\}$ be a set of features, with values in $\mathbb{R}$. Let A and C be two disjoint subsets of F, that is, $A \subset F$, $C \subset F$, and $A \cap C = \emptyset$. A QAR is a rule $X \Rightarrow Y$, in which features in A belong to the antecedent X, and features in C belong to the consequent Y, such that X and Y are formed by a conjunction of multiple boolean expressions of the form $F_i \in [v_1, v_2]$. The consequent Y is usually a single expression. In this proposal, QAR are used because the domain is a continuous domain.

It is important measure the quality of the rule in order to select the best rules and evaluate the results obtained by the proposed algorithm. In the ARs mining process, probability-based measures that evaluate the generality and reliability of ARs have been selected [18][19]. In particular, support is used to represent the generality of the rule and confidence, lift and leverage are used to represent the reliability of the rule. Others popular measures are conviction, gain, certainty factor and accuracy.

In most cases, it is sufficient to focus on a combination of support, confidence, and either lift or leverage to quantitatively measure the ”quality” of the rule. However, the real value of a rule, in terms of usefulness and actionability is subjective and depends heavily of the particular domain and business objectives.

III. Methodology

In this section we describes the main features of the proposed algorithm in order to discover ARs from datasets whose attribute are real data.

A. Search of Rules

In a continuous domain, it is necessary to group certain sets of values that share same features and therefore it is required to express the membership of the values to each group. Adaptive intervals instead of fixed ranges have been chosen to represent the membership of such values in this work. The search for the most appropriate intervals has been carried out by means of the proposed algorithm. Thus, the intervals are adjusted to find QAR with high values for support and confidence, together with other measures used in order to quantify the quality of the rule.

Our proposal is based on the NSGA-II approach [10], and its main purpose is to evolve the population based on the non-dominated sort of the solutions in fronts of dominance. The first front is composed of the non-dominated solutions of the population (the Pareto front), the second is composed of the solutions dominated by one solution, the third of solutions dominated by two, and so on. The operating scheme of the algorithm proposed can be seen in Figure 1. The overall complexity of the algorithm NSGA-II is $O(MN^2)$, which is governed by the nondominated sorting part of the algorithm.

In the population, each individual constitutes a rule. These rules are then subjected to an evolutionary process, in which the mutation and crossover operators are applied and, at the end of the process the best individual the Pareto front is designated as the best rule. Our proposal performs an IRL process (Iterative Rule Learning) [21] to penalize instances already covered by rules found by the algorithm, in order to emphasize the covering of instances still not covered. The IRL affects the generation of initial population in each evolutionary process which is described in Subsection III-C.

In order to optimize the mining of AR by the proposed algorithm, thus, rules with high quality and precision, two interestingness measures are selected as objectives:

- *Confidence*($X \implies Y$)[18]: Confidence is defined as the probability that instances satisfying X, also satisfy Y. In other words, it is the support of the rule divided by the support of the antecedent.

$$Conf(X \implies Y) = P(X \mid Y) = \frac{sup(X \implies Y)}{sup(X)} \quad (1)$$

  where $sup(X)$ is the support of the antecedent that is defined as the ratio of instances in the dataset that satisfy the antecedent X, and $sup(X \implies Y)$ is the

```
Inputs: Maximum number of rules ($MaxNumRules$),
Maximum number of generations ($MaxNumGen$)
Output: Population of the last generation

Multi-objective Algorithm(MaxNumRules, MaxNumGen)

Initialize the rule counter $r = 0$
Repeat
  1) Initialize the generation counter $t = 0$
  2) Initialize parent population $P_{t=0}$ based on instances covered by
     fewer rules.
  3) Evaluate the individuals of $P_{t=0}$ based on the measures selected
     as objectives.
  4) $P_{t=0}$ is ranked using the Fast non dominated Sort [10] that
     consists in sorting the individuals of a population in different
     Pareto fronts ($F$) according to their non dominance.
     Repeat
       a) an offspring population $Q_t$ of same size as $P_t$ is generated
          using crossover and mutation operators over the individuals
          of $P_t$ selected using binary Tournament selection-based
          method [20]
       b) The individuals of $P_t$ and $Q_t$ are merged into $R_t$ and the
          Fast Non dominated Sort is carried out.
       c) The next population $P_{t+1}$ consists of the $N$ best individuals
          of $R_t$.
          Initialize the front counter $i = 0$.
          Repeat
              If the current level of $R_t$ ($F_i$, $i-th$ Pareto front) has
              less than or equal to $N$ individuals, the individuals
              of $F_i$ are added to the population $P_{t+1}$.
              In other case,
                  if the current level of $R_t$ ($F_i$, $i-th$ Pareto
                  front) has more than $N$ individuals, the best
                  individuals are used to fill the population of
                  next generation ($P_{t+1}$), and for that purpose,
                  the Crowding distance assignment [10]is
                  used in order to sort the population of the
                  current level and select the best individuals
                  that represent the best rules.

              Increment the front counterr ($i = i + 1$)

          While the next population $P_{t+1}$ is not complete.
       d) Increment the generation counter ($t = t + 1$)
     While the maximum number of generations is not reached.
  5) Return best individual, thus, the rule in the first Pareto front
     ($F_1$) which reach a higher crowding distance value.
  6) Penalize the instances covered by the best rule found.
  7) Increment the rule counter ($r = r + 1$)
While the number of desired rules is not reached.
Return the best rules found.
```

Figure 1. General scheme of the algorithm.

support of the rule, thus, the percentage of instances in the dataset that satisfy X and Y simultaneously.

- *Leverage*($X \Longrightarrow Y$)[19]: Leverage measures the proportion of additional cases covered by both X and Y above those expected if X and Y were independent of each other. Leverage takes values inside [-1, 1]. Values equal or under value 0, indicate a strong independence between antecedent and consequent. On the other hand values near 1 are expected for an important association rule. Values above 0 are desirable. In addition, leverage is a lower bound for support, and therefore, optimizing only the leverage guarantees a certain minimum support (contrary to optimizing only the confidence or only the lift).

  $$Lev(X \Longrightarrow Y) = sup(X \Longrightarrow Y) - sup(X)sup(Y) \quad (2)$$

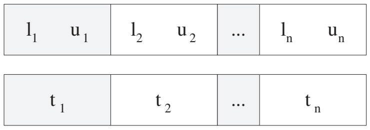
  where $sup(Y)$ is the support of the consequent of the rule, that is, the ratio of instances in the dataset that satisfy the consequent Y.

  The proposed algorithm doesn't use a threshold for minimum support and minimum confidence.

The different parts of the algorithm are defined in the following subsections.

B. Individuals Codification

The lower and upper limits of the intervals of each attribute will be represented by the different genes of an individual. Because the attributes are continuous, individuals are represented by a real coding. An individual consists of a not fixed number of attributes less than n, which represents the number of attribute in the database. The representation of an individual consists in two data structures as shown in Figure 2. The upper structure includes all the attributes of the database, where l_j is the lower limit of the range and u_j is the upper limit. The bottom structure indicates the membership of an attribute to the rule represented by an individual. The type of each attribute t_j, can have three values: 0 when the attribute does not belong to the rule, 1 if it belongs to the antecedent of the rule and 2 when it belongs to the consequent part. If an attribute is wanted to be retrieved for a specific rule, it can be done by modifying the value equal to 0 of the type by a value equal to 1 o or 2 depending on the antecedent or consequent.



l_1 u_1	l_2 u_2	...	l_n u_n

t_1	t_2	...	t_n

Figure 2. Representation of an individual of the population.

C. Initial Population

The generation of the initial population in the proposed algorithm was carried out at the beginning of each evolutionary process and is perform such at least one chosen sample or instance of the dataset was covered. The samples of the dataset are selected based on their level of hierarchy. The hierarchy is organized according to the number of rules which cover a sample. Thus, the records are sorted by the number of rules that are covered and the samples covered by a few rules have a higher priority.

A sample is selected according to the inverse of the number of rules which cover such sample. Intuitively, the process is similar to roulette selection method where the parents are selected depending on their fitness. Thus, the samples covered by a few rules have a greater portion of roulette and, therefore, they will be more likely selected. In the first evolutionary process, all samples have the same probability to be selected. Constraints to generate individuals are given by the number of attributes that belong to rule represented by an individual, the number of attributes in the antecedents and consequents and the structure of the rule (attributes fixed or not fixed in consequent).

D. Genetic Operators

The genetic operators implemented in the genetic algorithm proposed are Crossover and Mutation described in [15]. In addition, a new Mutation operator has been added. Concretely, the Antecedent $\Longleftrightarrow$ Consequent Mutation that works as follow: If the type t_i of the selected attribute is antecedent (1), changed to consequent (2), else if the type t_i of the selected attribute is consequent (2), changed to antecedent (1).

IV. Results

We applied our methodology to the microarray datasets of Spellman and Cho for the budding yeast (Saccharomyces cerevisiae) cell-cycle [22] and [23]. These data were synchronized by three different methods: cdc15, cdc28, and alpha-factors. Therefore, these three gene expression data sets may be defined as statistically independent [24].

The same training experiments with cdc15 dataset used by Soinov et al. in [24] were analyzed to achieve a comparison between the two methods. We considered a set of well-described genes, which encode proteins important for cell-cycle regulation. We selected these genes for the performance analysis of the proposed method in order to establish comparisons with the previous study [24].

A. Parameters configuration

As the proposed algorithm is non-deterministic, it has been executed five times for the dataset. The main parameters are as follows: 100 for the number of the rules to obtain, 50 for the size of the population, 50 for the number of generations, 0.1 for the mutation probability p_{Mut} of the individuals, 0.2 for the mutation probability p_{MutGen} of each gene in the individual.

B. Discussion of Results

In order to choose the best individual (rule) of each generation, the individual with the highest support value in the first Pareto front has been selected in order to cover the maximum number of examples by the obtained rules. We have extracted the relationships between attributes belonging to the antecedent and attributes belonging to the consequent for each AR found by the proposed algorithm in each run. For example, if we have the following rule:

$$A \in [0.2, 1.3] \Longrightarrow B \in [0.3, 1.2] \wedge C \in [0.5, 1.9]$$

the relationships or associations between the attributes of the antecedent and consequent of the rule are:

$$A \Longrightarrow B \text{ and } A \Longrightarrow C$$

Then, we have built a graph with associations derived from the rules, where each attribute that belongs to the rule is a graph node and each association obtained between attributes is an edge of the graph.

For the resulting graph, we performed the intersection between the graphs obtained in each of the five executions carried out by the algorithm in order to find the frequent interrelations between genes.

Table I shows some of the QAR obtained by the algorithm resulting after performing the intersection of the graphs constructed for each algorithm execution. The *Sup. Rule* column, shows the support of the rule that is the percentage of samples covered by the rule. The *Conf* column indicates the probability that instances satisfying the antecedent, also satisfy the consequent. The *Lev* column presents the leverage of the rule and measures the proportion of additional cases covered by both antecedent and consequent above those expected if they were independent of each other. The *Acc* column describes the accuracy of the rule and means the percentage success of the rule. The *CF* column presents the Certainty Factor of the rule. The interest of the rule is shown in column *Lift* and the *Amp* column presents the average amplitude of the intervals of the attributes belonging to each rule. It is important that the values of all interestingness measures of the AR are as high as possible.

For better understanding, Table I shows rules containing 2 attributes, one attribute in the antecedent and one in the consequent. Rules formed by 3 attributes are shown only for the relationships of genes that are not obtained in any rule of 2 attributes. Because the format of the rules obtained by the algorithm is not fixed, that is, any attribute may belong to the antecedent or the consequent, rules have been obtained with the same attributes but the sense of the implication of the association is different. For example, rules 0 and 1, rules 3 and 4, which are represented as directed edges in the graph in Figure 3.

We can see that the support value of all rules, between 25 % and 50 %, is good enough for the problem at hand. Equally remarkable, the values of confidence, certainty factor and accuracy for most of the rules is equal to 1 or very close to 1, which means that these measures have their highest value and indicates that the rule is totally accurate and the implication of the rule is perfect. The lift and leverage values are quite high, and this means that the rules are interesting and provides valuable information about antecedent and consequent occurring together in the dataset. In addition, the proportion of instances covered by both antecedent and consequent is greater than ones covered by antecedent and consequent separately. Leverage is a lower bound for support, so optimizing leverage guarantees

Table I
QUANTITATIVE ASSOCIATION RULES AND GENE-GENE ASSOCIATIONS INFERRED BY THE PROPOSED ALGORITHM

ID	Rule	Sup. Rule	Conf	Lev	Acc	CF	Lift	Amp	Gene-Gene associations inferred by our method	Soinov
0	$CLN1 \in$ [0.23, 1.21] $\Longrightarrow CLN2 \in$ [0.84, 1.72]	0.292	1	0.207	1	1	3.429	0.26	CLN1 CLN2	
1	$CLN2 \in$ [0.61, 1.72] $\Longrightarrow CLN1 \in$ [0.2, 1.21]	0.333	1	0.222	1	1	3	0.296	CLN2 CLN1	√
2	$CDC20 \in$ [-0.23, 0.91] $\Longrightarrow CLN1 \in$ [-1.34, -0.28]	0.5	0.857	0.184	0.875	0.688	1.582	0.332	CDC20 CLN1	√
3	$CLB1 \in$ [-1.37, -0.23] $\Longrightarrow CLB2 \in$ [-1.74, -0.07]	0.5	1	0.25	1	1	2	0.496	CLB1 CLB2	√
4	$CLB2 \in$ [-1.74, -0.15] $\Longrightarrow CLB1 \in$ [-1.37, -0.23]	0.458	1	0.229	0.958	1	2	0.483	CLB2 CLB1	√
5	$CLB6 \in$ [-0.92, 0.09] $\Longrightarrow CLB5 \in$ [-0.58, 0.02]	0.375	1	0.219	0.958	1	2.4	0.285	CLB6 CLB5	√
6	$CLB5 \in$ [-0.58, -0.11] $\Longrightarrow CLB6 \in$ [-0.92, 0.09]	0.333	1	0.208	0.958	1	2.667	0.254	CLB5 CLB6	√
7	$CLN2 \in$ [0.61, 1.72] $\Longrightarrow CLB5 \in$ [0.25, 1.08]	0.333	1	0.222	1	1	3	0.352	CLN2 CLB5	
8	$CLB2 \in$ [0.42, 1.24] $\Longrightarrow CLB5 \in$ [-1.02, 0.08]	0.458	1	0.172	0.833	1	1.6	0.399	CLB2 CLB5	
9	$CLB2 \in$ [-0.24, 0.93] $\Longrightarrow SW15 \in$ [-0.56, 0.75]	0.542	1	0.226	0.958	1	1.714	0.418	CLB2 SW15	√
10	$CDC34 \in$ [-1.17, 0.06] $\Longrightarrow MBP1 \in$ [0.28, 1.27]	0.458	1	0.248	1	1	2.182	0.45	CDC34 MBP1	√
11	$MBP1 \in$ [0.52, 1.27] $\Longrightarrow CDC34 \in$ [-1.17, -0.19]	0.417	1	0.243	1	1	2.4	0.352	MBP1 CDC34	√
12	$MBP1 \in$ [0.52, 1.13] $\Longrightarrow SKP1 \in$ [-1.47, -0.13]	0.375	1	0.203	0.917	1	2.182	0.358	MBP1 SKP1	√
13	$SKP1 \in$ [-0.83, -0.24] $\Longrightarrow MBP1 \in$ [0.52, 1.27]	0.33	1	0.194	0.917	1	2.4	0.241	SKP1 MBP1	
14	$SW15 \in$ [0.3, 0.77] $\Longrightarrow CLN2 \in$ [-1.88, -0.14]	0.375	1	0.18	0.875	1	2	0.321	SW15 CLN2	√
15	$CLB1 \in$ [0.07, 1.27] $\Longrightarrow CLN2 \in$ [-1.88, -0.14]	0.458	0.917	0.208	0.917	0.833	1.833	0.469	CLB1 CLN2	
16	$CLB1 \in$ [-1.37, -0.53] $\Longrightarrow SW15 \in$ [-1.46, -0.34]	0.333	1	0.194	0.917	1	2.4	0.349	CLB1 SW15	√
17	$CLB2 \in$ [-1.74, -0.15] $\Longrightarrow$ $CLB1 \in$ [-1.37, -0.23] $\wedge CLN2 \in$ [0, 1.72]	0.458	1	0.248	1	1	2.182	0.481	CLB2 CLN2	√
18	$MBP1 \in$ [-1.12, 0] $\wedge$ $CDC53 \in$ [-0.62, 0.09] $\Longrightarrow$ $SKP1 \in$ [-0.21, 0.74]	0.458	1	0.21	0.917	1	1.846	0.387	CDC53 SKP1	
19	$SW14 \in$ [-0.14, 0.3] $\wedge$ $CLB4 \in$ [0.09, 0.95] $\Longrightarrow$ $CDC34 \in$ [0.06, 0.68] $\wedge CLN1 \in$ [-0.59, 0.99]	0.417	1	0.243	1	1	2.4	0.387	SW14 CDC34	

a certain minimum support (contrary to optimizing only confidence or only lift).

C. Biological Relevance

The associations inferred by our approach are summarized in the tenth column of Table I. The eleventh column of Table I indicates gene-gene associations that were also inferred by the proposed methods by Soinov in [24] using the same dataset. The Gene Regulatory Network corresponding to the rules inferred by our approach and Soinov is shown in Figure 3 and 4, respectively.

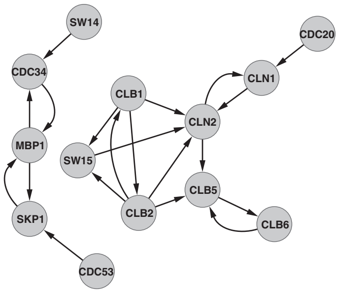


Figure 3. Directed graph obtained by the proposed algorithm.

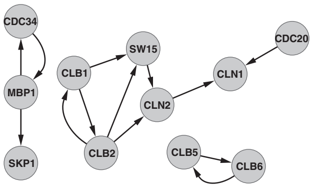


Figure 4. Directed graph obtained by Soinov.

In summary, all rules inferred by the decision-tree-based method [24] (13 in total) were also inferred by our approach, with the addition of new seven rules inferred only by our proposal. The biological relevance of the rules inferred by our approach was verified by analyzing whether such rules reflect functional properties relating to the different cell-cycle phase. The rules which are supported by the literature are: 3, 4, 5, 6, 9, 10, 11, 12, 14, 16. The rules 1 and 2 are consistent with the prior knowledge and are detected by Soinov. The rules which are not supported by the literature, i.e. 0, y and 7 are new hypothesis to analyze in the laboratory.

V. CONCLUSION

A multi-objective evolutionary algorithm for mining quantitative association rules has been proposed in this work. The approach is based on the well-known NSGA-II and has determined the intervals that form the rules without discretizing the attributes as a first step of the process. In order to evaluate its performance, the approach has been applied in a dataset and compared to other published results. The results report the relevance and significance in the group of genes found in the rules obtained for the problem studied in terms of support, confidence, accuracy, interest and leverage.

As a conclusion, an advantage of network reconstruction using our approach is that the method is able to construct a network correctly, i.e. reproducing the logic of a network consistent with the data as [24]. The network reconstructed from cell cycle yeast dataset is consistent with the knowledge store in the literature. Furthermore, the method can be improve by adding prior knowledge and more gene expression profiles. Our method constitute an interactive expert system for gene association networks, where the expert decides when to stop adding new gene expression profiles and what biological meaning represent the network.

Acknowledgment

The financial support from the Spanish Ministry of Science and Technology, project TIN2007-68084-C-00, and from the Junta de Andaluca, project P07-TIC-02611, is acknowledged.

References

[1] P. Brown and D. Botstein, "Exploring the new world of the genome with dna microarrays," *Nature Genet.*, vol. 21, no. Suppl., pp. 33–37, 1999.

[2] F. Azuaje and Y. D. adn DR Wagner, "Coordinated modular functionality and prognostic potential of a heart failure biomarker-driven interaction network," *BMC Syst. Biol.*, vol. 4, p. 60, 2010.

[3] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proceedings of the International Conference on Very Large Databases*, 1994, pp. 478–499.

[4] M. Vannucci and V. Colla, "Meaningful discretization of continuous features for association rules mining by means of a som," in *Proceedings of the European Symposium on Artificial Neural Networks*, 2004, pp. 489–494.

[5] E. D. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Publishing Company, 1989.

[6] J. Alcalá-Fdez, N. Flugy-Pape, A. Bonarini, and F. Herrera, "Analysis of the effectiveness of the genetic algorithms based on extraction of association rules," *Fundamenta Informaticae*, vol. 98, no. 1, pp. 1001–1014, 2010.

[7] J. Mata, J. L. Álvarez, and J. C. Riquelme, "Discovering numeric association rules via evolutionary algorithm," *Lecture Notes in Artificial Intelligence*, vol. 2336, pp. 40–51, 2002.

[8] X. Yan, C. Zhang, and S. Zhang, "Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support," *Expert Systems with Applications: An International Journal*, vol. 36, no. 2, pp. 3066–3076, 2009.

[9] K. Deb, *Multi-Objective Optimization Using Evolutionary Algorithms*. John Wiley & Sons, Inc., 2001.

[10] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: Nsga-ii," *Evolutionary Computation, IEEE Transactions on*, vol. 6, no. 2, pp. 182 –197, 2002.

[11] E. Zitzler, M. Laumanns, and L. Thiele, "Spea2: Improving the strength pareto evolutionary algorithm," *EUROGEN*, vol. 3242, no. 103, pp. 95 – 100, 2001.

[12] B. Alatas, E. Akin, and A. Karci, "MODENAR: Multi-objective differential evolution algorithm for mining numeric association rules," *Applied Soft Computing*, vol. 8, no. 1, pp. 646–656, 2008.

[13] H. Qodmanan, M. Nasiri, and B. Minaei-Bidgoli, "Multi objective association rule mining with genetic algorithm without specifying minimum support and minimum confidence," *Expert Systems with Applications*, vol. 38, no. 1, pp. 288–298, 2011.

[14] M. Martínez-Ballesteros, F. Martínez-Álvarez, A. Troncoso, and J. C. Riquelme, "Mining quantitative association rules based on evoluationary computation and its application to atmospheric pollution," *Integrated Computer-Aided Engineering*, vol. 17, pp. 227–242, 2010.

[15] M. Martínez-Ballesteros, F. Martínez-Álvarez, A. Troncoso, and J. Riquelme, "An evolutionary algorithm to discover quantitative association rules in multidimensional time series," *Soft Computing*, vol. 15, no. 10, pp. 2065–2084, 2011.

[16] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2006.

[17] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," in *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, 1993, pp. 207–216.

[18] L. Geng and H. Hamilton, "Interestingness measures for data mining: A survey," *ACM Comput. Surv.*, vol. 38, no. 3, p. 9, 2006.

[19] G. Piatetsky-Shapiro, "Discovery, analysis and presentation of strong rules," in *Knowledge Discovery in Databases*, 1991, pp. 229–248.

[20] B. Miller and D. Goldberg, "Genetic algorithms, tournament selection, and the effects of noise," *Complex Systems*, vol. 9, pp. 193–212, 1995.

[21] G. Venturini, "SIA: A Supervised Inductive Algorithm with genetic search for learning attribute based concepts," in *Proceedings of the European Conference on Machine Learning*, 1993, pp. 280–296.

[22] P. Spellman, G. Sherlock, M. Zhang, V. Iyer, K. Anders, M. Eisen, P. Brown, D. Botstein, and B. Futcher, "Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization," *Mol Biol Cell 1998,*, vol. 9, pp. 3273–3297, 1998.

[23] R. Cho, M. Campbell, E. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. Wolfsberg, A. Gabrielian, D. Landsman, D. Lockhart, and R. Davis, "A genome-wide transcriptional analysis of the mitotic cell cycle," *Mol Cell*, vol. 2, pp. 65–73, 1998.

[24] L. Soinov, M. Krestyaninova, and A. Brazma, "Towards reconstruction of gene networks from expression data by supervised learning," *Genome Biology*, vol. 4, p. R6, 2003.